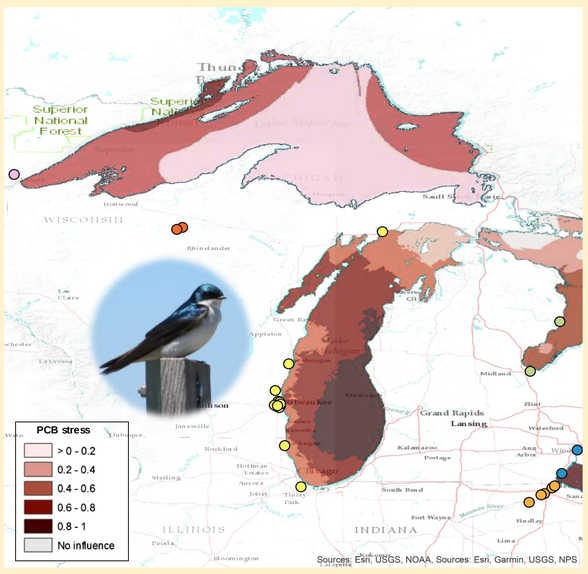
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Population Genetics of Tree Swallows, in Collaboration with NCGAS

Sheri Sanders[1], Charles M. Mansfield[2], Bhavya Papudeshi[1], Carrie Ganote[1], Chi Yen Tseng[2], Thomas W. Custer[3], Christine M. Custer[3], Cole W. Matson[1], Tom Doak[1]
National Center for Genome Analysis Support, Indiana University[1], Baylor University, CRASR and Environmental Science[2], United States Geological Survey[3]

Abstract (REVISED)
The National Center for Genome Analysis Support (NCGAS) provides training and computational resources in an effort to train biologists to approach historically-difficult, non-model problems with large biological data sets. For example, our collaborators at Baylor University work with Tree Swallow (*Tachycineta bicolor)*, using RNAseq data in population genetics and toxicology. Working with the NCGAS, they assembled a de novo transcriptome assembly for the Tree Swallow, for which there is no genome. Variant calling using the transcriptome identified 66,169 single nucleotide polymorphisms (SNPs) across 144 samples. They were then able to identify phylogeographic structuring across the Great Lakes Region, including accurate grouping populations distributed across smaller geographic scales (e.g. along the Maumee River). SNPs were also used to assess population heterozygosity and genetic diversity. This project required large scale data handling, large memory machines to assembly the transcriptome, and advanced Linux skills to manage the data and analyses. NCGAS provided the computation resources and training on the Linux environment and data management. Further assistance was provided in consultation and problem solving - leading to a high level of independence and competency of the graduate student researcher.

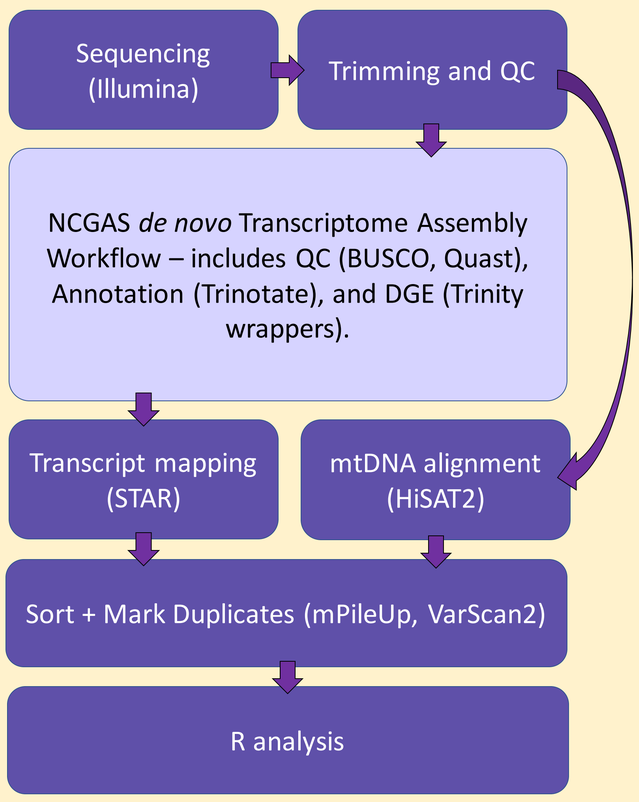
Who is NCGAS?
The mission of the NCGAS is to enable the biological research community of the US to analyze, understand, and make use of the vast amount of genomic information now available. We also support field stations.



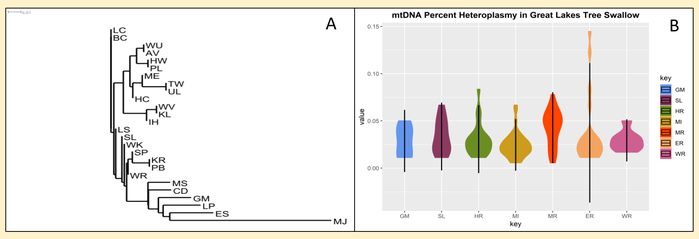
Introduction
- RNAseq analysis is often limited to differential gene expression analysis and pathway analysis, but population genetics information can be mined – allowing for future toxicology analysis across populations.
- *de novo* transcriptome assembly allows analyses on non-model species. There is no nuclear genome, but SNPs can be identified within transcripts. Reads were also mapped to the published mitochondrial genome for analysis.
- Concentrations of PCBs, PAHs, and other contaminants have been documented for years in the Great Lakes Region, which will be investigated in reference to population genetics.
- 204 Tree Swallows were sampled over 4 years. Samples were sequenced with 2x150bp Illumina and analyzed on NCGAS resources.

Computational Resources
This study benefitted from free access to:
- IU's Carbonate – our general use genomics machine
- PSC's Bridges – 12TB memory cluster for science use

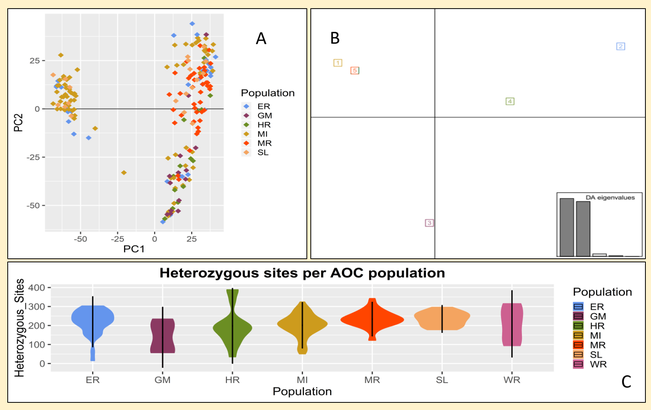
In addition to in-house compute resources, NCGAS has free allocations on the national compute and storage system, XSEDE. We are Domain Champions for both genomics and field stations, and work closely with Indiana University's Jetstream team and Pittsburg Supercomputing Center's Bridges Team to provide software and tools.



Methods and Results
- The NCGAS *de novo* transcriptome assembly pipeline was used to create the reference transcriptome. Mapping and clean up of the data was performed on NCGAS hardware, with consultation by NCGAS staff.
- Populations did not sort by location – they may be confounded by small sample size and contamination profiles, which is being addressed currently.
- However, population genomic analyses were proven viable with RNAseq data, allowing for preliminary measures of heterozygosity and heteroplasmy.



Mitochondrial Analysis: A) Phylogeny using genetic distance calculated from Weirs F_{ST}. B) Percent mtDNA heteroplasmy.



Transcriptome Analysis:
A) Principal Component Analysis of 85,689 SNPs in the transcriptome. Highly contaminated Lake Michigan, River Raisin and Maumee River form distinct clusters from other areas of concern.
B) Discriminant Analysis splits 204 samples into 5 theoretical populations.
C) Transcript heterozygosity between areas of concern.

NCGAS Training
Our collaborators took part in our training workshops. We offer a variety of tutorials, guides, and workshops that help in similar analyses:
- *de novo* transcriptome assembly workshop (3 days)
- Six-part R class, focusing on reading the language, visualization, and flexibility.
- Active blog on new tools, resources, and common questions.
- Materials for getting started on Unix, building genome browsers, planning a genome project, using the free cloud, building automatic reporting into field sensors, and more.

Conclusions
- RNAseq can be used in population genomics studies to analyze population structures.
- mtDNA information can be extracted from RNAseq projects, allowing the use of heteroplasmy as a proxy for mutation rates.
- Completion of analysis will help our collaborators identify SNPs that are driving population structure and associated with toxicology measures.
- Sample collection will continue throughout the Great Lakes Region, allowing for higher resolution of structure and diversity.
- Collaboration between NCGAS and Baylor University continues, allowing for additional training of graduate students, as well as quick access to new tools.

Want to work with NCGAS?
Want to explore microbiomes, but aren't quite sure how much data you need to sequence or what analyses to use? Planning a genome assembly but not sure which technologies best fig your needs? Need a bioinformatician that understands your data? We're happy to collaborate on long term projects as contributing authors – half of our staff have PhDs in Biology!

Other projects we have collaborated on include population phylogeography of strawberry dart frogs, microbiome differences in kelp forests and coral reefs, and genetic diversity of rattlesnakes.

Contact us at help@ncgas.org
/ncgasiu @ncgas ncgas.org